Evaluating Models

Metrics for model performance

```
augment(tree_fit, new_data = frog_test) %>%
  metrics(latency, .pred)
#> # A tibble: 3 × 3
#>   .metric .estimator .estimate
#>   <chr>   <chr>          <dbl>
#> 1 rmse    standard      59.2
#> 2 rsq     standard       0.380
#> 3 mae     standard      40.2
```

- RMSE: difference between the predicted and observed values ↓
- R^2: squared correlation between the predicted and observed values ↑
- MAE: similar to RMSE, but mean absolute error ↓

Metrics for model performance

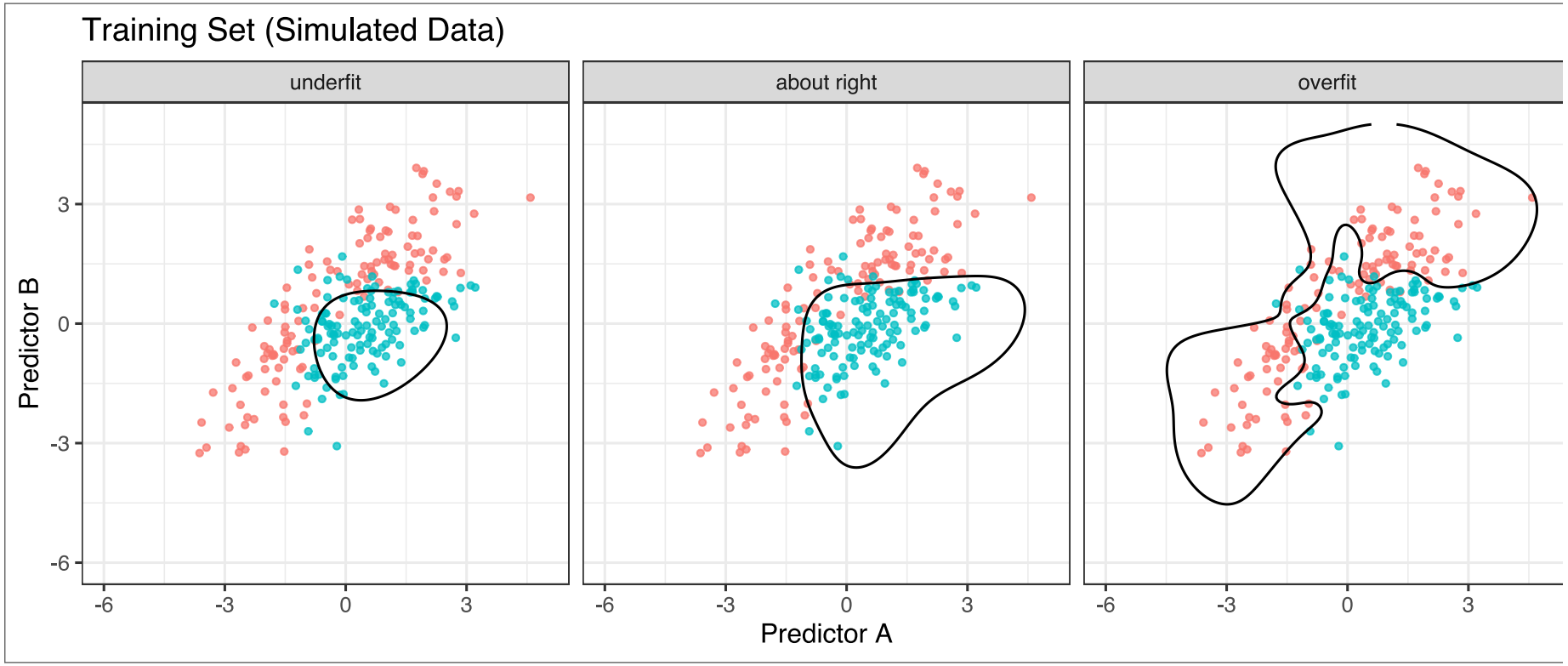
```
augment(tree_fit, new_data = frog_test) %>%
  rmse(latency, .pred)
#> # A tibble: 1 × 3
#>   .metric .estimator .estimate
#>   <chr>   <chr>          <dbl>
#> 1 rmse    standard        59.2
```

Metrics for model performance

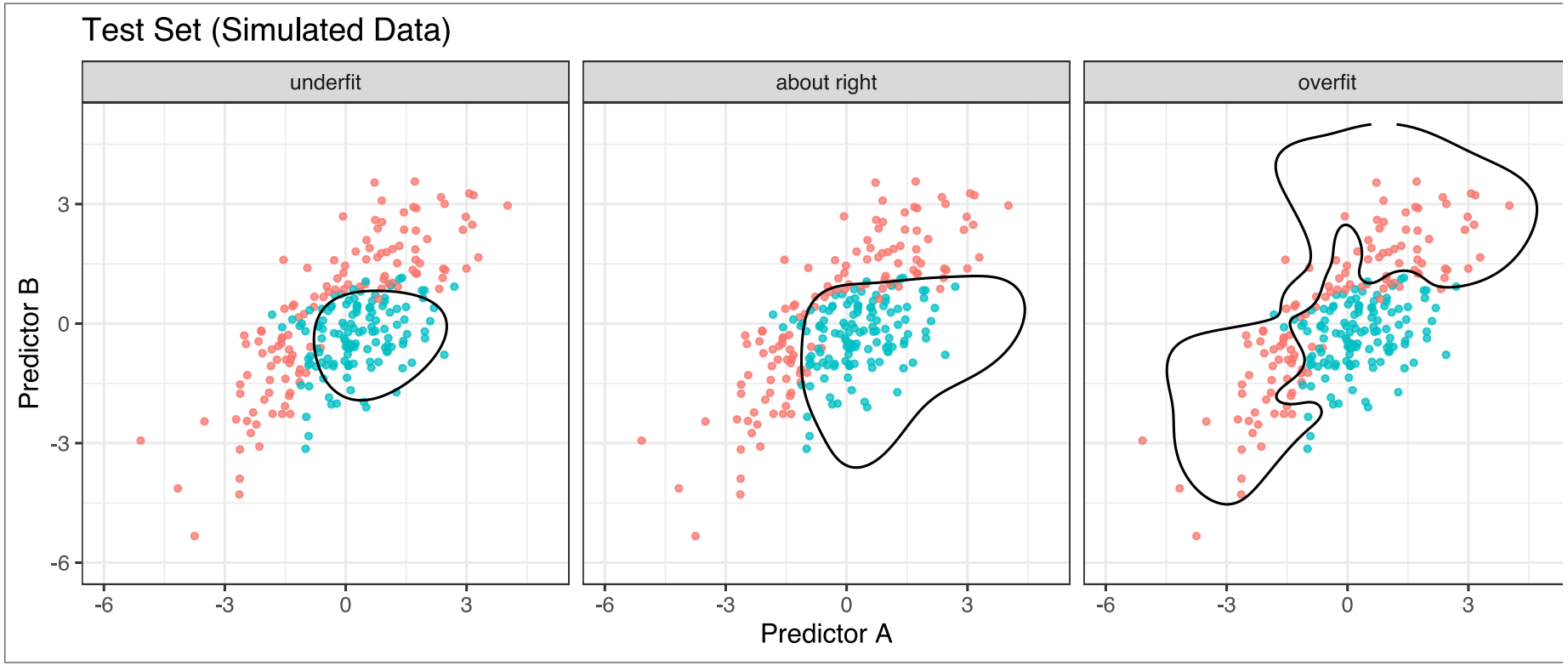
```
augment(tree_fit, new_data = frog_test) %>%
  group_by(reflex) %>%
  rmse(latency, .pred)
#> # A tibble: 3 × 4
#>   reflex .metric .estimator .estimate
#>   <fct>  <chr>   <chr>          <dbl>
#> 1 low    rmse    standard        94.3
#> 2 mid    rmse    standard       101.
#> 3 full   rmse    standard        51.2
```

⚠ DANGERS OF OVERFITTING ⚠

Dangers of overfitting ⚠



Dangers of overfitting ⚠

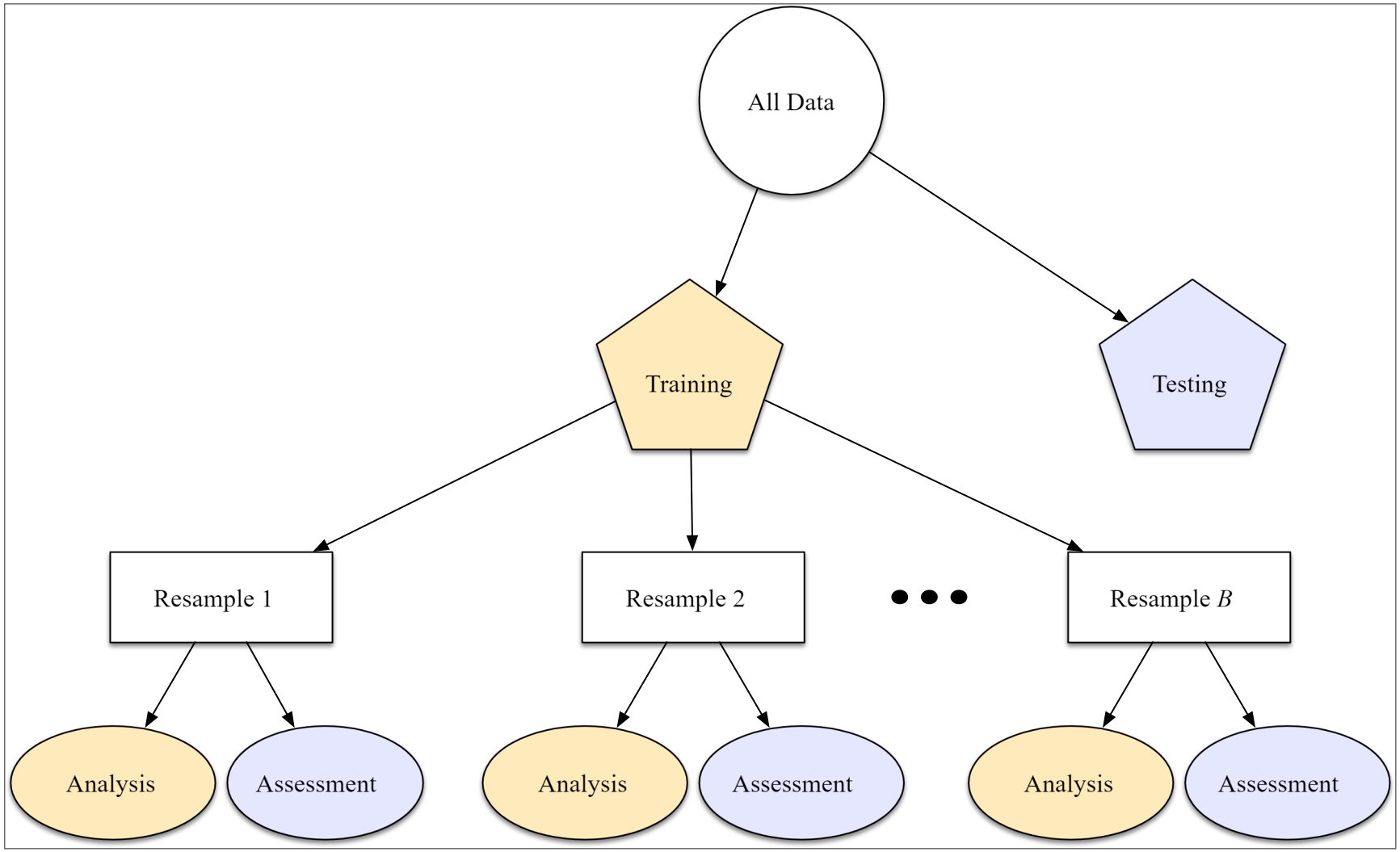


We call this “resubstitution” or “repredicting the training set”
What if we want to compare more models?
And/or more model configurations?
And we want to understand if these are important differences?

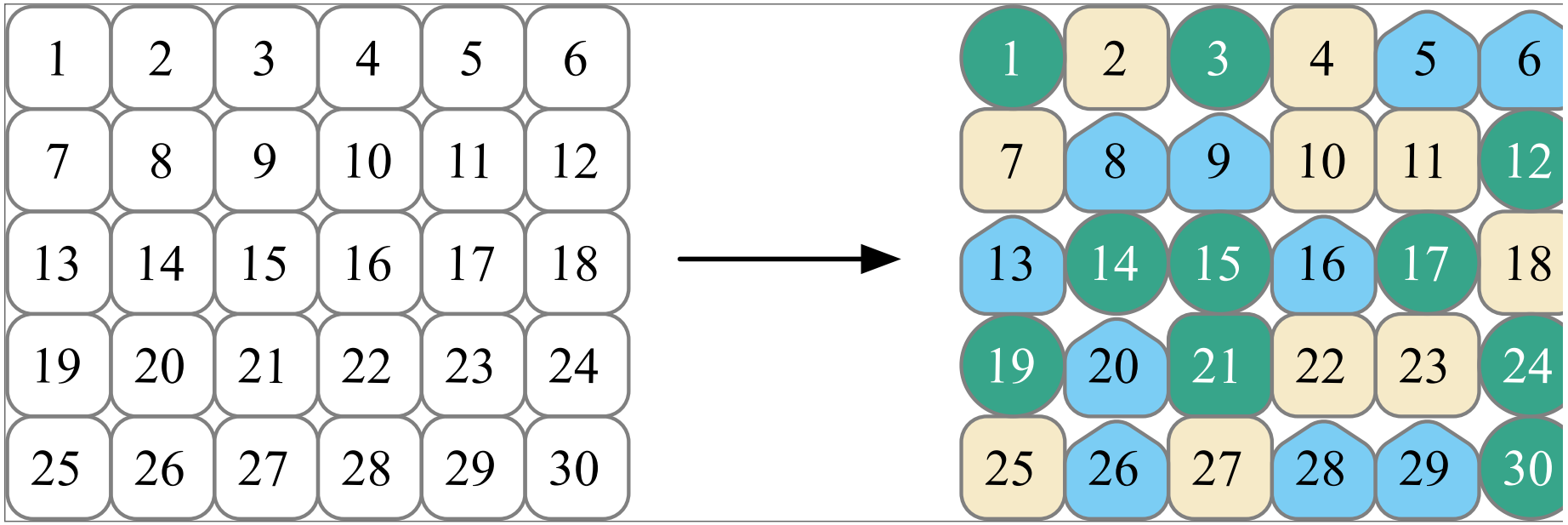
The testing data are precious 💎

How can we use the *training* data to compare and evaluate different models? 🤔

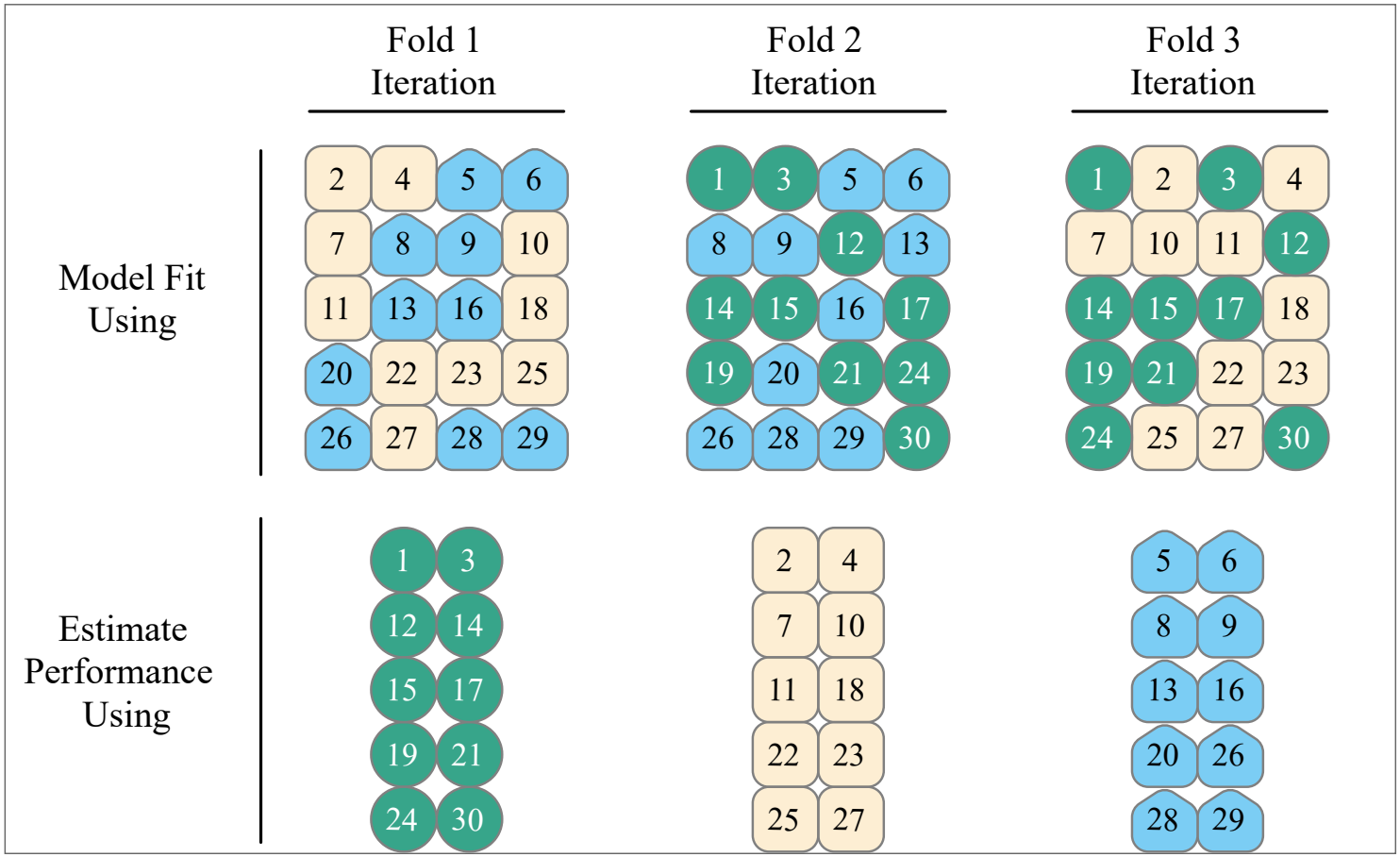
Resampling


All Data
Training
Testing
Resample 1
Resample 2
Resample B
Analysis
Assessment
Analysis
Assessment
Analysis
Assessment

Cross-validation

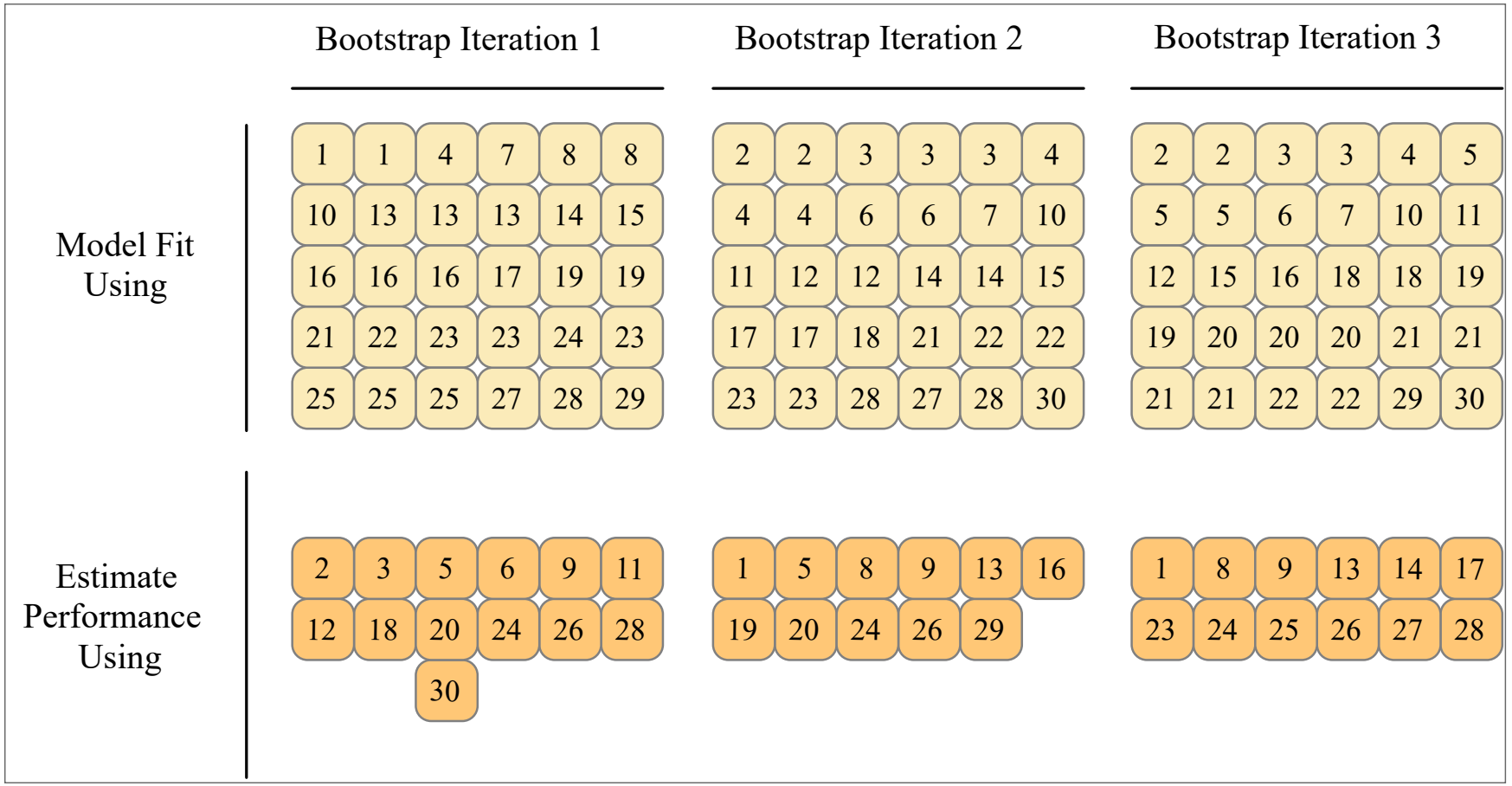


Cross-validation



Alternate resampling schemes

Bootstrapping



Error